DOCUMENT RESUME

ED 249 266 TM 840 619

AUTHOR Becker, Betsy Jane
TITLE Power Differences among Tests of Combined Significance.
PUB DATE Apr 84
NOTE 21p.; Paper presented at the Annual Meeting of the American Educational Research Association (68th, New Orleans, LA, April 23-27, 1984).
PUB TYPE Speeches/Conference Papers (150) -- Reports - Research/Technical (143)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS Effect Size; Hypothesis Testing; *Meta Analysis; Research Methodology; Sample Size; *Statistical Analysis; *Statistical Significance
IDENTIFIERS *Power (Statistics)

ABSTRACT

Power is an indicator of the ability of a statistical analysis to detect a phenomenon that does in fact exist. The issue of power is crucial for social science research because sample size, effects, and relationships studied tend to be small and the power of a study relates directly to the size of the effect of interest and the sample size. Quantitative synthesis methods can provide ways to overcome the problem of low power by combining the results of many studies. In the study at hand, large-sample (approximate) normal distribution theory for the non-null density of the individual p value is used to obtain power functions for significance value summaries. Three p-value summary methods are examined. Tippett's counting method, Fisher's inverse chi-square summary, and the logit method. Results for pairs of studies and for a set of five studies are reported. They indicate that the choice of a "most-powerful" summary will depend on the number of studies to be summarized, the sizes of the effects in the populations studied, and the sizes of the samples chosen from those populations. (BW)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED249266

Power Differences Among Tests of Combined Significance

Betsy Jane Becker

Michigan State University

U.S. DEPARTMENT OF EDUCATION
NATIONAL INSTITUTE OF EDUCATION
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☒ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official NIE position or policy.

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

B J Becker

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

TM 840 619

Paper presented at
the annual meeting of the
American Educational Research Association

New Orleans, April 1984

Power Differences Among Tests of Combined Significance

The quantitative review of research

The review of research has long been important for the development of theory in the social sciences. Feldman (1971), Pillemer and Light (1980), and Taveggia (1974) have discussed the functions of research cumulation and its role in scientific inquiry. Until recently the narrative or discursive review of studies has been the most common method of summarizing studies in social science.

An alternative which has been suggested to the traditional narrative literature review is the quantitative cumulation of study findings. Some quantitative techniques for combining study results have been available for many years. Rosenthal (1978) has reviewed a group of methods with a long history of inquiry: methods for summarizing one-sided significance values (levels) or p values. These are the methods I will examine below.

One reason that quantitative methods for summarizing research are important especially for the social sciences relates to the adequacy of individual studies. The fact that researchers call for and are willing to perform research replications indicates that they are aware of the fallibility of single studies. No reasonable researcher expects to obtain the final, unique and exact answer in his or her study. The use of statistical inference in individual studies or experiments assumes that the results obtained are probabilistic; researchers implicitly acknowledge that their conclusions may be wrong.

Power is an indicator of the ability of a study, or specifically a statistical analysis, to detect a phenomenon (e.g., an effective treatment, a relationship between two variables) that does in fact exist. Power is a

theoretical characteristic, since if the existence of a phenomenon could be demonstrated or were known statistical inference would be unnecessary.

Several people (e.g., Brewer, 1972; Cohen, 1962) have shown that the power of much social science research is low. That is, most single studies do not have great ability to detect the phenomena of interest to the researcher. (Apparently the researcher does well not to put great faith in the results of these single studies!) The issue of power is especially crucial for social science research because the numbers of subjects (sample sizes) and the effects and relationships studied tend to be small, and the power of a study relates directly to the size of the effect of interest and the sample size. Quantitative synthesis methods can provide ways to overcome the problem of low power by combining the results of many studies.

Using Quantitative Summary Methods

If we are going to summarize a series of studies we should first decide what hypotheses we have or what questions we would like to ask about the studies. We can then choose from the available summaries the particular method that will give us the most information. Specifically we will choose the method that is suitable for our data (in the present case, p values), that tests our hypothesis, and that is best at providing information about the kinds of outcomes that interest us.

The Tests of Combined Significance

Though there are many methods for summarizing significance values they are all derived from a few general principles. Rosenthal (1978) presents eight methods and there are many more. The names of the methods are presented in Table 1; summaries derived by a common approach are classed together. (See Rosenthal's article for more detail about tests of combined significance.)

Power of Tests of Combined Significance

Numerous articles in the statistical literature concerning the properties of p-value summaries indicate that there is not a single "best" method. One much investigated property of the summaries has been power, which is the focus of this paper as well.

As noted above, the power of a statistical method is the (conditional) probability that the method suggests rejection of the null hypothesis, given that some alternative hypothesis truly describes the population being examined. The power of a method will vary across different alternative hypotheses.

When methods for analyzing data from one group are examined, alternative hypotheses can be easily described by the size of the difference between the suggested (null) effect and the true effect. For example, the size of the population mean value (μ) would represent the size of the true effect if the null hypothesis were that $\mu = 0$ (e.g., that there was no effect for that group). However, the hypotheses underlying tests of combined significance represent several groups (studies), thus there are many possible deviations from the null condition, which is that all populations studied have the same true null hypothesis. Becker (1982) has discussed the hypotheses which can be examined by one-sided tests of combined significance (i.e., p-value summaries).

For tests of combined significance, then, power will vary according to the outcomes found in the series of studies. When the result of a series of studies is not null, any number of studies in the series may be of populations with non-null results. Also the size of the effect of interest may vary across studies for which the effect is non-null. In some cases, studies with large effects contrary to the effect of interest may be in a series for which a significance value summary indicates support for the effect in the direction of interest. Sometimes mixed and even conflicting

results are plausible or anticipated. For example, a new curriculum may be more effective than an old one if implemented by trained teachers, but less effective otherwise.

The numerous and diverse outcomes for series of studies make it difficult to conceptualize and to calculate power for significance value summaries. Ironically, information about the power of tests of combined significance is especially requisite because of this variety of outcomes.

Examining the Power of Tests of Combined Significance

Past Studies of Power

Since the p-value summary methods vary widely in form, it is not surprising that they vary also in their power to detect different outcomes. Some authors have studied the power of these methods. Several examinations of power of these so-called omnibus tests are found in the statistical literature, including work by George (1977), Koziol and Perlman (1978), Pearson (1938), Zelen and Joel (1959), and others. In the social science literature, Hedges and Olkin (1981) discussed power for the synthesis procedures known as vote-counting, which are related to the methods I have classed as "counting" methods in Table 1.

For most outcomes the summary methods do not behave uniformly. Some methods (e.g., Tippett's method) are most sensitive to the situation in which one study is of a population with a particularly large effect and the others show small (or null) effects. The same methods may be relatively poor at detecting consistent but only moderate-sized effects for sets of studies.

One aspect of power that has thus far been largely ignored is performance of the summaries under the case of conflicting outcomes. It is more often the case in social science that series of studies will not completely agree than that they will agree totally. Occasionally the

results will even conflict (see, e.g., reviews of studies of open education by Giaconia and Hedges (1982), and of sex differences in influenceability by Becker (1983)).

The Fisher method has been noted to lend support to (i.e., is relatively powerful at detecting) two opposing alternative hypotheses (Adcock, 1954), while inspection of some of the others (e.g., the Stouffer and logit methods) indicates that they would allow conflicting results to cancel out and would indicate support for the null hypothesis. If we were studying a phenomenon which was likely to show conflicting outcomes, we would be wise not to use a summary that had low power for detecting such outcomes. Were we to use such a summary, we would be more likely to (wrongly) conclude that the studies all were of populations which had null effects than if we used a more powerful summary.

How Power is Studied

Statisticians have used several approaches in examining power for tests of combined significance, including, for example, the enumeration of all possible outcomes for a combination of discrete test statistics (Lancaster, 1949), and the simulation of study outcomes (Koziol & Perlman, 1978). I will take a different approach, utilizing large-sample (approximate) normal distribution theory for the non-null density of the individual p value (Lambert, 1978) to obtain power functions for the summaries.

For the purposes of this paper I will present results for only a subset of the p-value summary methods. The methods that will be examined are the Tippett (1931) method, the Fisher (1932) method, and the logit method (George, 1977). The methods in this subset have properties that are representative of those for the full set of methods, and will adequately illustrate the importance of considering possible outcomes and power differences before applying a significance value summary. Table 2 gives

formulas and decision rules for these summary methods.

Lambert's Result

Lambert (1978) and Lambert and Hall (1982) have presented a large-sample distribution for the logarithmically transformed (single) p value. Lambert presented the result in the context of examining the efficiency of the test statistic with which the p is associated, but conveniently it can be used to obtain distributions for functions of several p's as well.

The large-sample distribution given by Lambert is

$$n^{-1/2}\left(-\frac{1}{n}\log(p) + c(\theta)\right) \sim N(0, \tau^2(\theta))$$

where θ is the population value of the effect size parameter, $c(\theta)$ is the exact slope of the test statistic from which p arose, $\tau^2(\theta)$ is the variance of the transformed p (which relates to $c(\theta)$) and n represents the sample size or degrees of freedom for the test statistic. Formulas for the exact slope (or convergence rate) $c(\theta)$ and for $\tau^2(\theta)$ for most common test statistics have been derived (through investigations of test efficiency).

Lambert's (1978) work indicated that this lognormal distribution was most accurate for large non-centrality parameters (large values of $\sqrt{n}\,\theta$). Lambert suggested replacing log (p) with log ($\sqrt{n}p$) in the above formula to increase distribution accuracy. Further study (Note 1) indicates that even this form is inaccurate for θ values near zero ($\theta \leq 0.15$) for most reasonable sample sizes.

Deriving Distributions for the Summaries

If we have k independent log-transformed p values we can use Lambert's result presented in a multivariate format to describe their distribution. Furthermore, if we have functions of sets of p values we can apply convergence theorems to obtain distributions for the functions. Convergence theorems are methods used by mathematical statisticians to obtain

distributions for new variables. One standard procedure is called Cramer's delta method (Rao, 1973), which under many conditions allows one to obtain the distribution of a continuous function of a variable that has a known distribution. Though the delta method was not directly applicable to Lambert's result, an alternate convergence theorem (Note 2) proved useful.

Thus using these "tools" we can obtain distributions for the p-value summaries, which are simply functions of k independent p's. From these distributions (which will be normal in large samples) we can obtain power functions and make comparisons among the summaries. By varying the values of the θ's and n's we can examine different kinds of outcomes. The procedures used and the outcomes examined in the power study are described below.

Calculation of the Power Functions

The calculation of power functions for tests of combined significance is in general difficult because of the intractable form of the non-null distribution of the individual p value. Under the null hypothesis all p values are uniformly distributed between 0 and 1. The non-null distribution of the p value depends on both the form of the statistic from which the p value is drawn and its parameters, as well as the size of the sample for which the statistic is calculated.

Thus under each alternative hypothesis the distribution for a test of combined significance is not a combination of a set of identical distributions. This is true even when the p's combined are from the same kinds of test statistics (e.g., t tests). The distribution for a p-value summary is even more complex when p's from different test statistics are combined. In terms of Lambert's result above, we would have different values and different functional forms for $c(\theta)$ and $\tau^2(\theta)$ for the k tests.

For the computations in this paper we examined the power of a (hypothetical) combination of one-sample t tests. We varied the sizes of the true population effects and the sizes of the samples for which the t

tests would have been calculated. Thus our values for c and τ varied but the forms of their functions were the same.

The Power Study

To examine differences between the summaries in their likelihoods to detect different possible outcomes we must calculate their power. We first outline a set of hypothetical outcomes for a series of studies. There are obviously many from which to select. For each outcome and each summary, we then calculate the probability that the null hypothesis for the series will be rejected at a particular significance level.

Below we detail the hypothetical outcomes for which power calculations were made. The series lengths (k) and samples sizes (n_i) were selected on the basis of an examination of 1,018 sample sizes from 105 series of studies from 29 research reviews in Psychological Bulletin and Review of Educational Research in 1980 through 1983 (Note 1).

Over sixty percent of the 105 series comprised less than 10 studies, and the median k value was 8. Surprisingly $k = 2$ (i.e., pairs of studies) was a modal value for the distribution: 12 percent of the series were pairs of studies. In this paper we will compute power for sets of two and five studies.

Pairs of studies. The mean study sample sizes for 9 pairs of studies (in the 105 mentioned above) were $n_1 = n_2 = 56$. The values $n_1 = n_2 = 50$ were used to represent the equal-sized-studies case. To examine the effect of imbalance in sample size we examine the case in which one study has a sample size of 25 and the other has a sample size of 75.

For each sample-size pair we examine the case in which the first study is from a population with a moderate positive effect (e.g., a moderate mean or correlation value) and the population effect for the other study varies. Because of inaccuracy of the large-sample distribution for the log(p),

outcomes involving small effects were not examined.

Set of five studies. The mean study sample size for 6 series (from the above-mentioned 105 series) of 5 studies was $n_i = 64$. To be a bit conservative we will examine the case in which each of the 5 studies has 50 people. The unbalanced case we will investigate is that in which two studies have n's of 25, one has an n of 50, and the rest have n's of 75.

For each sample-size configuration we examined the situation in which one population effect varied and the others were moderate in size.

Computation

Computations were performed via FORTRAN programs utilizing calls to the International Mathematical and Statistical Libraries cumulative inverse normal distribution subroutine MDNOR. Programs were written by the author.

Results

Pairs of studies. The sizes of the hypothesized population effects in the studies to be summarized relate directly to power for the tests of combined significance. Tables 3 and 4 present power values for the Fisher, logit, and Tippett summaries for .05 significance level tests on pairs of studies.

The power values increase steadily for all three tests, though at slightly different rates. For large values of θ_2, the tests have comparable power. When θ_2 is moderate to small, however, there are considerable differences among the three methods. For instance, for balanced sample sizes (Table 3) the power of the Tippett method exceeds that of the logit by about 15 percent when $\theta_2 = 0.15$. The Fisher method's power exceeds that of the Tippett summary by as much as 36 percent when θ_2 is 0.20 or 0.25. (There is a small chance that these differences are exaggerated because of differential dependence of the power functions on the Lambert approximation.)

In Table 4 the same trends are present, but the power values therein are also influenced by the sample size imbalance. When the smaller study is of a population with a smaller effect θ_2 the power of all methods is higher than in the balanced n's case for the small θ_2 values. Conversely, power is lower for the high θ_2 values because here the smaller study represents the population with the larger effect. The large study ($n_1 = 75$) influences the power more than the small study and thus it increases power only minimally because θ_1 is small.

Figure 1 presents an alternative view of the power of two of the methods that reveals the complexity of the power results. The fifty percent equi-power contours represent the pairs of θ values which produce power values of 0.5 for the Fisher and Logit summaries. Effects represented by θ values lying between the curves and the axes are likely to be detected less than 50 percent of the time. Points beyond the equi-power curves represent outcomes that would be likely to be detected more than 50 percent of the time.

The figure illustrates that in cases where one study has a small effect, the Fisher method achieves 50 percent power for a smaller-sized effect in the second population than does the logit method. In the range where both studies are of populations that have moderate sized effects, the logit method achieves .50 power for smaller effects than does the Fisher method. This suggests that if we are interested in detecting moderate and equal-sized effects in two populations, the logit method is more likely to do so than the Fisher method.

Sets of five studies. Table 5 gives power values for the combination of five studies for the Fisher and logit summaries. Trends similar to the ones seen in Tables 3 and 4 are present in this table as well. However, in the combinations of 5 studies the power of the Fisher method is considerably

higher than that of the logit method for all outcomes examined. This agrees with the work of most statisticians, who have found that the power of the Fisher method is adequate for most purposes.

Summary

The results presented here illustrate the power of only three of a set of nearly twenty ways to summarize p values. (Further work to investigate the power of the other methods is in progress.) Yet even with this small illustration we can see that the chioce of a "most-powerful" summary will depend on the number of studies to be summarized, the sizes of the effects that we expect to find in the populations studied, and the sizes of the samples that we choose from those populations.

The choice of a best significance value summary might be based of the calculation of power for situations that represent expectable outcomes for the series of studies to be examined. Representative effect size values would have to be chosen on the basis of theory about the substantive effects being investigated. These values could be used with the actual sample size values for the accumulated studies. The power functions derived for the summaries would give values that could be compared, and the summary with the highest power for the expectable outcomes could be used.

Reference Notes

1. Becker, B.J. (1984). Accuracy of Lambert's log-normal distribution for the p value. Unpublished manuscript, Michigan State University, East Lansing, MI.

2. Becker, B.J. (1983). Addendum to "The Analysis of Significance Values". Addendum to dissertation proposal, The University of Chicago, Chicago, IL.

References

Adcock, C.J. (1960). A note on combining probabilities. Psychometrika, 25, 303-305.

Becker, B.J. (1982). Conceptual problems in the summarization of significance values. Paper presented at the annual meeting of the American Educational Research Association, New York.

Becker, B.J. (1983). Influence again: A comparison of methods for meta-analysis. Paper presented at the annual meeting of the American Educational Research Association, Montreal, Canada.

Brewer, J.K. (1972). On the power of statistical tests in the American Educational Research Journal, American Educational Research Journal, 9, 391-401.

Cohen, J. (1962). The statistical power of abnormal-social psychological research: A review. Journal of Abnormal and Social Psychology, 65, 145-153.

Feldman, K.A. (1971). Using the work of others: Some observations on reviewing and integrating. Sociology of Education, 44, 86-102.

Fisher, R.A. (1932). Statistical methods for research workers. (4th ed.), London: Oliver and Boyd.

George, E.O. (1977). Combining independent one-sided and two-sided statistical tests: Some theory and applications. Unpublished doctoral dissertation, The University of Rochester.

Giaconia, R.M. & Hedges, L.V. (1982). Identifying features of effective open education. Review of Educational Research, 52, 579-602.

Hedges, L.V. and Olkin, I. Vote-counting methods in research synthesis. Psychological Bulletin, 1980, 88, 359-369.

Koziol, J.A. and Perlman, M.D. (1978). Combining independent chi-squared tests. Journal of the American Statistical Association, 73, 753-763.

Lambert, D.M. (1978). P-values: Asymptotics and robustness. Unpublished doctoral dissertation, The University of Rochester.

Lambert, D. and Hall, W.J. (1982). Asymptotic lognormality of p-values. The Annals of Statistics, 10, 44-64.

Lancaster, H.O. (1949). The combination of probabilities arising from data in discrete distributions. Biometrika, 36, 370-382.

Pearson, E.S. (1938). The probability integral transformation for testing goodness of fit and combining independent tests of significance. Biometrika, 30, 134-148.

Pillemer, D.B. and Light, R.J. (1980). Synthesizing outcomes: How to use research evidence from many studies. Harvard Educational Review, 50, 176-195.

Rao, C.R. (1973). Linear statistical inference and its applications. New York: John Wiley & Sons, Inc.

Rosenthal, R. (1978). Combining results of independent studies. Psychological Bulletin, 85, 185-193.

Taveggia, T.C. (1974). Resolving research controversy through empirical cumulation. Sociological Methods and Research, 2, 395-407.

Tippett, L.H.C. (1931). The methods of statistics (1st ed.). London: Williams and Norgate, Ltd.

Zelen, M. and Joel, L. (1959). The weighted compounding of two independent significance tests. Annals of Mathematical Statistics, 30, 885-895.

Table 1

Methods for Summarizing Significance Values

Uniform Distribution Theory Methods

- Counting Methods
  - Wilkinson
  - Tippett
  - Sign Test
  - Chi-square
- Linear Combinations
  - Sum of P's
  - Mean P

Probability Transformation Methods

- Inverse Normal Summaries
  - Stouffer
  - Weighted Stouffer
  - Mean Z
- Inverse t Summaries
  - Inverse t (Winer)
  - Weighted Inverse t
- Inverse Chi-square Summaries
  - Inverse Chi-square
  - Weighted Inverse Chi-square
  - Fisher
  - Weighted Fisher (Good)
- Logistic Summaries
  - Logit
  - Weighted Logit

Table 2

Formulas and Decision Rules for Three Methods

Method	Decision Rule: Reject H_0 if:
Fisher	$-2 \sum_{i=1}^{k} \log(p_i) > \chi^2(2k, \alpha_*)$
Logit	$-\left(\frac{3\pi^2}{k}\right)^{-\frac{1}{2}} \sum_{i=1}^{k} \log\left[\frac{p_i}{1-p_i}\right] > \underline{t}(5k+4, \alpha_*)$
Tippett	$\min(p_1, \ldots, p_k) < 1 - (1-\alpha_*)^{1/k}$

Note: α_* is the pre-set series-wise type I error level.

Table 3

Power for Fisher, Logit, and Tippett Summaries

Nominal significance level is .05

Combinations of two studies, $\theta_1 = 0.25$

Sample sizes are $n_1 = 50$ and $n_2 = 50$

θ_2	Fisher	Logit	Tippett
0.15	.097	.089	.102
0.20	.167	.151	.123
0.25	.260	.237	.191
0.30	.369	.340	.300
0.35	.482	.452	.428
0.40	.591	.563	.555
0.45	.689	.665	.669
0.50	.773	.753	.763
0.55	.840	.825	.838
0.60	.892	.881	.894
0.65	.930	.922	.933
0.70	.956	.951	.960
0.75	.974	.971	.977
0.80	.985	.983	.987
0.85	.992	.991	.993
0.90	.996	.995	.997
0.95	.998	.998	.998
1.00	.999	.999	.999

Table 4

Power for Fisher, Logit, and Tippett Summaries

Nominal significance level is .05

Combinations of two studies, $\theta_1 = 0.25$

Sample sizes are $n_1 = 75$ and $n_2 = 25$

θ_2	Fisher	Logit	Tippett
0.15	.176	.159	.249
0.20	.213	.193	.249
0.25	.260	.237	.254
0.30	.315	.289	.273
0.35	.377	.348	.311
0.40	.441	.411	.365
0.45	.507	.477	.431
0.50	.572	.543	.503
0.55	.635	.607	.575
0.60	.693	.668	.645
0.65	.746	.724	.710
0.70	.794	.774	.768
0.75	.836	.819	.818
0.80	.871	.857	.861
0.85	.900	.889	.895
0.90	.925	.915	.923
0.95	.944	.936	.945
1.00	.959	.953	.961

Table 5

Power for Fisher and Logit Summaries

Combinations of five studies, $\theta_1 = \ldots = \theta_4 = 0.25$

Nominal significance level is .05

	Sample Sizes			
	$n_1 = \ldots n_5 = 50$		$n_1=50, n_2=n_3=75, n_4=n_5=25$	
θ_5	Fisher	Logit	Fisher	Logit
0.15	.469	.237	.523	.287
0.20	.517	.281	.545	.310
0.25	.572	.338	.576	.338
0.30	.632	.406	.603	.372
0.35	.694	.480	.637	.410
0.40	.752	.558	.672	.453
0.45	.806	.636	.708	.497
0.50	.853	.709	.743	.544
0.55	.893	.775	.777	.591
0.60	.924	.832	.806	.637
0.65	.948	.878	.839	.682
0.70	.966	.915	.866	.725
0.75	.978	.943	.890	.765
0.80	.987	.963	.910	.801
0.85	.992	.976	.928	.834
0.90	.995	.986	.943	.863
0.95	.998	.992	.956	.889
1.00	.999	.995	.966	.910

FIGURE: VALUES OF θ_1 AND θ_2 FOR WHICH THE FISHER AND LOGIT SUMMARIES HAVE POWER = 0.5

P VALUES ARE FROM ONE-SAMPLE T TESTS WITH $N_1 = N_2 = 50$